DOCUMENT RESUME

ED 185 038 TM 800 078

AUTHOR Downing, Steven M.; Maatsch, Jack L.
TITLE The Effects of Clinically Relevant Multiple-Choice Items on the Statistical Discrimination of Physician Clinical Competence.
INSTITUTION Michigan State Univ., East Lansing.
SPONS AGENCY National Center for Health Services Research (DHEW/PHS), Hyattsville, Md.
PUB DATE [78]
GRANT HS02038
NOTE 8p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Clinical Experience; *Competence; *Difficulty Level; Graduate Medical Students; Higher Education; Medical Students; *Multiple Choice Tests; Physicians; *Test Items; Test Reliability; *Test Validity
IDENTIFIERS Emergency Medical Service; *Emergency Medicine Examination

ABSTRACT

To test the effect of clinically relevant multiple-choice item content on the validity of statistical discriminations of physicians' clinical competence, data were collected from a field test of the Emergency Medicine Examination, test items for the certification of specialists in emergency medicine. Two 91-item multiple-choice subscales were identified: medium difficulty (MD)--items of ideal difficulty for norm-referenced achievement tests, and high clinical-relevance (HCR)--items that correlated highly with an independent criterion measure of simulated clinical performance. A stepwise discriminant analysis of these two subscale scores showed that HCR was more statistically discriminating of residency-eligible physicians, second-year residents, and fourth-year medical students than was the MD scale. The HCR scale was significantly more reliable and significantly less difficult than the MD scale. It was concluded that: (1) multiple-choice examinations, calibrated to a simulated clinical performance measure, can be very valid discriminators of levels of clinical competence; and (2) selection of multiple-choice items for medium difficulty is less efficient than selection of items for their high correlation to a simulated clinical performance measure. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED185084

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ABSTRACT

THE EFFECTS OF CLINICALLY RELEVANT MULTIPLE-CHOICE ITEMS ON THE STATISTICAL DISCRIMINATION OF PHYSICIAN CLINICAL COMPETENCE

Steven M. Downing, Ph.D.
Jack L. Maatsch, Ph.D.
Office of Medical Education Research and Development
Michigan State University

This research was carried out to test the effect of clinically relevant multiple-choice item content on the validity of statistical discriminations of physicians' clinical competence. Data were collected from a field test of the Emergency Medicine Examination, a library of items intended for the certification of specialists in Emergency Medicine.

Two 91-item multiple-choice subscales were identified: 1) Medium Difficulty (MD)--items of ideal difficulty for norm-referenced achievement tests, and; 2) High Clinical-Relevance (HCR)--items that correlated highly with an independent criterion measure of simulated clinical performance.

A Stepwise Discriminant Analysis of these two subscale scores showed that HCR was more statistically discriminating of residency-eligible physicians, second-year residents, and fourth-year medical students than was the MD scale.

Other psychometric characteristics of these two subscales were also investigated. The HCR scale was significantly more reliable ($r=.95$) than the MD scale ($r=.88$). Also, the HCR scale was significantly less difficult ($\bar{p}=.72$) than was the MD scale ($\bar{p}=.60$).

It was concluded that the relevance of multiple-choice item content to clinical medicine is important to the discriminant validity of objective examinations intended to certify medical specialists. Practically, these results suggest to certifying boards that: 1) Multiple-choice examinations, calibrated to a simulated clinical performance measure, can be very valid discriminators of levels of clinical competence, and; 2) Selection of multiple-choice items for medium difficulty may not be the most efficient strategy to use to produce the most valid and reliable measures of a certified specialist's clinical competence.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Steven M. Downing
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800 078

THE EFFECTS OF CLINICALLY RELEVANT MULTIPLE-CHOICE ITEMS ON THE STATISTICAL DISCRIMINATION OF PHYSICIAN CLINICAL COMPETENCE

Steven M. Downing, Ph.D.
Jack L. Maatsch, Ph.D.
Office of Medical Education Research and Development
Michigan State University[1]

Purpose of This Study

The major purpose of this study was to investigate the effect of clinically relevant multiple-choice item content on the validity of discriminations of physician clinical competence. Research data were collected during a field test of the Emergency Medicine Examination (EME)--a multi-format examination ultimately intended for certification of specialists in Emergency Medicine.

The study compared the relative discriminating power of a multiple-choice subscale composed of items selected for high clinical-relevance to a subscale of items selected for middle difficulty. High clinical-relevance items were selected for their highest correlation with an independent rating of simulated clinical performance, while medium difficulty items were selected for their middle difficulty levels. Additionally, the internal-consistency reliabilities and the mean item difficulties of these two multiple-choice subscales were investigated.

Review of the Literature

During the past decade, increasing pressure has been brought on specialty certifying boards to demonstrate the validity of their measurement procedures for the prediction of the clinical competence of certified specialists. Several empirical studies, investigating the relationship between scores on certification examinations and measures of physician performance, have been reviewed by Williamson (1976). One of the earliest studies, the Teamster Study (Trussell, 1962), concluded that the certification status of physicians was, for the most part, unrelated to the quality of health care delivered. This study was replicated and the findings confirmed by Morehead and others (1964).

McGuire and Williamson (1968) studied the performance of general practitioners, non-certified, and certified physicians on patient management problems. No statistical difference in performance was observed among these three groups. Pawluk and others (1976) report low positive or negative correlations between various certification format scores and an independent measure of quality of care for Canadian Family Practitioners. Gonnella (1973) found low correlations between scores on multiple-choice examination scores and measures of physician diagnostic accuracy and proper patient management.

---

[1]This research was supported in part by a grant from the National Center for Health Services Research (HS02038). The data, test formats, scoring mechanisms and all related examination development and validation procedures described in this paper were developed for the American College of Emergency Physicians. The American Board of Emergency Medicine, which will subsequently administer the first certification examination, reserves the right to use all or part of the test library and methodologies developed by the American College of Emergency Physicians. Reprints: Steven M. Downing, Ph.D., Michigan State University, the Office of Medical Education Research and Development, A-202 East Fee Hall, East Lansing, Michigan 48824.

Some of the best evidence for the lack of criterion-related validity for certification examinations was presented by Payne and Lyons (1972). From a large-scale study of the correlates to competent health-care delivery, these researchers concluded that board certification, type of specialization, years of practice, and hospital size did not correlate well with process audit ratings of physician performance.

On the other hand, Maatsch and others (1978) reported that both Multiple-Choice and Simulated Clinical Encounter scores are sensitive to years of training and experience in Emergency Medicine. Preliminary analyses of data from a field test of the EME showed that a total score on the objective format correlated .83 with independent ratings of clinical performance on realistic, patient-game simulations of typical emergency department clinical problems.

Methodology

This section presents a brief description of EME construction, the sample of subjects for this study, the experimental design, and the statistical methodology employed.

Instrument Development

The EME was developed by the American College of Emergency Physicians (ACEP) and the Office of Medical Education Research and Development (OMERAD), Michigan State University. This Examination consists of two objective-item formats (364 Pictorial Multiple-Choice and Multiple-Choice items), 12 Patient Management Problems (PMP), and 12 examiner-administered Simulated Clinical Encounters (SCE).

All Examination materials were field tested with ninety-four subjects who represented four distinct groups in training, experience, and ability to deliver health care in Emergency Medicine. These groups were: 1) Residency-Eligible Emergency Physicians: Subjects potentially eligible to take the EME by virtue of graduation from an approved residency in Emergency Medicine and at least one year of continuous practice in Emergency Medicine (n=22); 2) Practice-Eligible Emergency Physicians: Subjects potentially eligible for EME by virtue of five years of continuous practice in Emergency Medicine (n=14); 3) Emergency Medicine Residents: Residents beginning their second post-graduate year (n=36); and 4) Fourth-Year Medical Students: Volunteer subjects beginning the final year of pre-doctoral study (n=22).

Emergency physicians were selected by a national stratified random sampling procedure designed to select subjects judged by their peers as competent clinicians and who were, therefore, certifiable in Emergency Medicine. Residents in this study represented a national stratified random sample of all Emergency Medicine residency programs; subjects were selected for a wide-range of clinical competence. Medical students were paid volunteers.

All subjects completed all examination materials during two and one-half days of testing. Objective items were administered under standard testing conditions and the examiner-administered SCEs were presented and rated by well trained emergency physicians examiners. The interclass inter-rater reliability of the SCE grand mean rating was .79 (Maatsch et al., 1978). Group membership of individual subjects was unknown to SCE examiners.

Research Hypothesis

The major research hypothesis under investigation here is that an empirically defined multiple-choice subscale, consisting of high clinical-relevance items, will statistically separate residency-eligible physicians, residents, and students better than a multiple-choice subscale consisting of medium difficulty items.

The following subscales were selected for this investigation: 1) Medium Difficulty Subscale (MD): 91 items with ideal item p-values (proportion correct) for norm-referenced achievement tests (Ebel, 1972). P-values ranged from p=.50 to .69 with mean p=.603. All items had positive point-biserial item-total score discrimination indices; and 2) High Clinical-Relevance Subscale (HCR): 91 items with the highest item-mean SCE rating point-biserial correlations. Item-criterion correlations ranged from r=.33 to .68 with the median r=.38. The mean p-value for this scale was .721.

Results

TABLE I
RAW-SCORE GROUP DISCRIMINATION:
MEDIUM DIFFICULTY VS HIGH CLINICAL-RELEVANCE

	Residency-Eligible		Residents		Students		Wilks'	$F_{2,77}$
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Lambda	
High Clinical-Relevance	80.23	7.67	70.81	7.59	42.55	9.98	.2353	125.21*
Medium Difficulty	66.77	8.67	56.83	7.89	39.77	7.14	.3686	65.94*

*$p \leq .0001$

Table I presents the raw-score group discrimination of three criterion groups by each subscale. This analysis shows that each subscale statistically separates groups, since scale means are ranked in the predicted direction and each F-ratio is statistically significant. Wilks' Lambda, a multivariate index of the group discriminating power of a variable (Tatsuoka, 1971), indicates that the HCR scale accounts for more discriminating power than the MD scale. (A small Lambda indicates high discrimination power.)

TABLE 2
SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS:
HIGH CLINICAL-RELEVANCE VS MEDIUM DIFFICULTY

Step Number	Scale Name	F to Enter	Wilks' Lambda
1	High Clinical Relevance	125.21	.235
2	Medium Difficulty	4.38	.211

Table 2 gives the results of a Stepwise Discriminant Analysis of these two scales for the three criterion groups in this study. This procedure selected

the most discriminating variable first, entered this variable into the analysis, and then chose the second-best discriminating variable for entry into the analysis. From Table 2 it is apparent that the HCR scale is considerably more discriminating of known groups than the MD scale, since there is little change in Lambda from step one to step two in this analysis.

Two discriminant functions were derived by the Stepwise Discriminant Analysis. In the first function--the HCR function--the standardized discriminant function coefficient (2.374) was about 6.7 times more important than MD (-.357) to the separation of known groups.

TABLE 3
CLASSIFICATION ANALYSIS USING HIGH
CLINICAL-RELEVANCE AND MEDIUM DIFFICULTY DISCRIMINANT FUNCTIONS

		Predicted[a]		
	N	Residency-Eligible	Residents	Students
Residency-Eligible	22	15(68.2)	7(31.8)	0(0)
Residents	36	6(16.7)	29(80.6)	1(2.8)
Students	22	0(0)	1(4.5)	21(95.5)

[a]Number in parentheses indicates percentage of classification for that group.

Table 3 presents a classification analysis of the predicted group membership versus the actual group membership, using the two discriminant functions extracted by this analysis. A total of 81.3 percent of subjects were correctly classified by both discriminant functions--HCR and MD. A chi-square test statistic (χ^2=82.66, df=4,$p<.0001$) was computed to test an hypothesis of chance classification accuracy. The high clinical-relevance scale, taken separately, classified 76.3 percent of subjects in known groups correctly, while the medium difficulty scale classified only 71.2 percent of cases correctly.

Subscale Reliability

Internal-consistency reliability was calculated for each 91-item subscale for the N=80 subjects in this study. The KR20 r for the HCR scale was .954, while r=.878 for the MD scale. An F-test of this observed difference in scale reliability is given, following Hoyt (1941) and shown by Downing (1979), by a ratio of the Mean Squares associated with the reliability calculation for each subscale. For this null hypothesis, the F-test (F=2.06, df=79, 79, $p<.05$) showed that the HCR subscale was statistically higher in internal-consistency reliability than the MD subscale.

Item Difficulty

TABLE 4
SUBSCALE MEAN ITEM DIFFICULTY
N=80

Subscale	Mean	Standard Deviation	Mean P-Value	Z-Value
High Clinical-Relevance	65.63	16.96	.721	
Medium Difficulty	54.86	12.88	.603	13.48*

*$p<.0001$

Table 4 gives the mean item difficulty for each scale. A correlated Z-test of the difference in scale means shows that the HCR scale is significantly less difficult than the MD scale.

Conclusions and Discussion

The following conclusions follow from the results of this study:

1. Multiple-choice items can be written and calibrated to be powerfully valid discriminators of varying levels of clinical competence.

2. Selection of multiple-choice items for their high correlation with a valid and reliable independent measure of clinical performance is more efficient for the valid discrimination of physician clinical competence than selection of items for a middle difficulty criterion.

3. Selection of multiple-choice items for their high correlation to a simulated clinical performance measure yields a more reliable measurement than selection of items for their middle difficulty.

4. Clinically relevant items are less difficult for clinicians than items typically used for certification or licensure.

These findings must be viewed as tentative and generalizations to other settings must be cautious, since subjects for this study may be atypical of subjects in other certification settings. However, the results of this study do suggest that the relevance of multiple-choice item content to the typical practice of clinical medicine is important to the valid statistical discrimination of levels of physician competence.

It is also interesting to note that high clinical-relevance items tend to be less difficult than typical, norm-referenced achievement items. Yet, the less difficult clinically relevant items yield highly reliable measurements of physician competence.

If these results generalize to other specialty certification settings, certifying boards would do well to select objective-item content that is maximally related to the typical and competent practice of clinical medicine.

Further research is indicated to replicate these findings, using larger samples of subjects. Validation and cross-validation studies are also essential. Another important area of research suggested by this study is the determination of the content selection rules for HCR scales.

References

Downing, S.M. "An Analysis of the Effects of Different Multiple-Choice Item Selection Strategies on the Reliability and Validity of Measures of Physician Competence in Specialty Certification," 1979, A Dissertation at Michigan State University.

Ebel, R.L. Essentials of Educational Measurement, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

"Emergency Medicine Condition/Skills List," Journal of the American College of Emergency Physicians, August, 1976.

Gonnella, J.S. "Evaluation of Patient Care: An Approach," Journal of the American Medical Association, Vol. 214, No. 11, 1973, pp. 2040-3043.

Hoyt, C.J. "Test Reliability Estimated by Analysis of Variance," in Principles of Educational and Psychological Measurement, eds. W.A. Mehrens and R.L. Ebel, Chicago, Illinois: Rand McNally and Company, 1967.

McGuire, C.; Williamson, J.W. "Consecutive Case Conference: An Educational Evaluation," Journal of the American Medical Association, Vol. 43, No. 10, 1968, pp. 1068-1074.

Maatsch, J.L.; et al. "Toward a Testable Theory of Physician Comeptence: An Experimental Analysis of a Criterion-Referenced Specialty Certification Test Library," in Proceedings of the Seventeenth Annual Conference on Research in Medical Education, Washington, D.C.: Association of American Medical Colleges, 1978.

Morehead, M.A.; Trussell, R.E.; Ehrlich, J. A Study of the Quality of Hospital Care Secured by a Sample of Teamster Families in New York City. New York: Columbia University Press, 1964.

Pawluk, W.; et al. "Concurrent Validity of the Canadian Certification Examination in Family Medicine," in Proceedings of the Fifteenth Annual Conference on Research in Medical Education, Washington, D.C.: Association of American Medical Colleges, 1976.

Payne, B.C.; Lyons, T.F. Method of Evaluation and Improving Personal Medical Care Quality, Office Care Study for Hawaii Medical Association, Ann Arbor, Michigan: University of Michigan School of Medicine, 1972.

Tatsuoka, M.M. Multivariate Analysis: Techniques for Educational and Psychological Research, New York: John Wiley and Sons, Inc., 1971.

Trussell, R.E. The Quantity, Quality and Costs of Medical and Hospital Care Secured by a Sample of Teamster Families in the New York Area, New York: Columbia University Press, 1962.

Williamson, J.W. "Validation by Performance Measures," in Conference on Extending the Validity of Certification, Chicago, Illinois: The American Board of Medical Specialties, 1976.